REMARKS

Applicants respectfully request that the above-identified application be re-examined.

The July 27, 2004, final Office Action ("Office Action") rejected Claims 1-4, 6, and 10-20 under 35 U.S.C. § 102(e) as being anticipated by the teachings of U.S. Patent No. 6,418,433 (Chakrabarti et al.). Claims 5 and 7-9 were rejected under 35 U.S.C. § 103(a) as being unpatentable in view of Chakrabarti et al. While applicants believe that the rejection is clearly in error, in order to advance the prosecution of this application, independent Claim 10 has been amended to include language previously added to independent Claim 1. The amendment has been made so that Claim 10 more particularly points out and distinctly claims the subject matter applicants regard as their invention.

Prior to discussing in detail why applicants believe that all of the claims in this application are allowable, a brief description of applicants' invention and a brief description of the teachings of the cited and applied reference are provided. The following discussions of applicants' invention and the cited and applied reference are not provided to define the scope or interpretation of any of the claims of this application. Instead, these discussions are provided to help the United States Patent and Trademark Office better appreciate important claim distinctions discussed thereafter.

Applicants' Invention

Applicants' invention is directed to an improved way of retrieving information pertaining to documents stored in a computer network. More specifically, the invention employs a **probabilistic model** to determine the likelihood that a document has changed since it was last accessed and, thus, to determine whether the document should be accessed during a current Web crawl. Preferably, the accuracy of the probabilistic model is continuously improved by training internal probability distributions to reflect the actual change rate pattern of the document to be accessed.

In one form, the invention is directed to a computer-implemented method for selectively accessing a document during a current crawl of a server computer, the document being identified by a document address specification and having been retrieved during a previous crawl. The method comprises determining whether to access the document during the current crawl **with the aid of a probabilistic model that is based on the probability that the document has changed since the previous crawl**. The method further comprises accessing the document if the determination produces an instruction indicative that the document at the document address specification should be accessed during the current crawl.

In another form, the invention is directed to a computer-readable medium having computer-executable instructions for retrieving one document in a plurality of documents from a remote server. When executed, the instructions: (i) maintain historical information associated with changes to the one document; (ii) initiate a crawl procedure for retrieving particular documents in the plurality of documents; and (iii) determine whether to access the one document from the remote server based on a **probabilistic analysis of the historical information** associated with the changes to the one document. **The probabilistic analysis of the historical information is based on the probability that the one document has changed since a previous crawl.**

It is clear from the foregoing summary that the present invention is based on a probabilistic analysis that is based on the probability that a document has changed since a previous crawl. Probability and probabilistic are statistical terms. The probability of an event occurring is the ratio of the number of times the event occurred in a larger number of trials. (*McGraw-Hill Dictionary of Scientific and Technical Terms*, 6th ed., page 1674.) In the present application, historical information is analyzed to determine probability.

<u>U.S. Patent No. 6,418,433 (Chakrabarti et al.)</u>

Chakrabarti et al. purportedly discloses a system and method for focused Web crawling. Chakrabarti et al.'s focused Web crawler learns to recognize Web pages that are relevant to the interest of one or more users from a set of examples provided by the users. Chakrabarti et al.'s focused Web Crawler explores the Web starting from the example set using statistics collected from the samples and other analyses of the link graph of a growing crawl database to guide itself toward relevant valuable resources and away from irrelevant and/or low-quality material. The Web crawler allegedly builds a comprehensive topic-specific library for the benefit of specific users. In summary, Chakrabarti et al. is directed to locating relevant documents. **Chakrabarti et al. is not directed to determining whether a document should be accessed during a current crawl based on the probability that the document has changed since a prior crawl.**

While Chakrabarti et al. purportedly does teach revisiting documents, **the basis for revisiting the documents is not based on a probabilistic model that the document has changed since a previous crawl.** Chakrabarti et al.'s basis for revisiting a document is based on the relevant priority of the document to the subject matter, not the probability that the document has been changed since the previous crawl. See Col. 10, lines 32-34, which states "the more relevant the page, the higher the priority for revisitation to check for subsequent changes." Lines 10-15 of Col. 10 of Chakrabarti et al. state that a page is classified as "good" in terms of relevancy using a topic analyzer.

Rejection of Claims 1-4, 6, and 10-20

As noted above, the Office Action rejected Claims 1-4, 6, and 10-20 under 35 U.S.C. § 102(e) as being fully anticipated by the teachings of Chakrabarti et al. The Office Action asserts that Chakrabarti et al. suggests each and every element of these claims. Applicants respectfully disagree. As described in more detail below, Chakrabarti et al. fails to disclose or suggest elements of both the independent and dependent claims of this application.

Independent Claim 1 reads as follows:

> 1. A computer-implemented method for selectively accessing a document during a current crawl of a server computer, the document being identified by a document address specification, the document having been retrieved during a previous crawl, the method comprising:
>
> determining whether to access the document during the current crawl with the aid of a probabilistic model that is based on the probability that the document has changed since the previous crawl; and
>
> accessing the document if the determination produces an instruction indicative that the document at the document address specification should be accessed during the current crawl.

The Office Action asserts that Col. 8, lines 53-67, of Chakrabarti et al. discloses "determining whether to access the document during the current crawl with the aid of a probabilistic model that is based on the probability that the document has changed since the previous crawl." Applicants respectfully disagree. Col. 8, lines 53-67, of Chakrabarti et al. reads as follows:

> On the other hand, when a worker thread is idle or near-idle, the logic moves to block 70 **to determine how many new Web pages** (that is, pages associated with outlinks in the link table 34) **to evaluate and how many old Web pages** (that is, pages already listed by URL in the Web page table 32) **to evaluate for potential changes to the old pages that might have occurred since the last time the old pages were considered by the system 10. The numbers of new and old pages generated at block 70 depends on, e.g., the amount of old pages already listed in the crawl database 30.** For example, if the database 30 is relatively full, more old pages will be checked for changes than new pages considered, whereas when the crawl database is relatively empty, more new pages will be evaluated than old pages checked for changes. (Emphasis added.)

Applicants submit that the foregoing language has nothing whatsoever to do with determining whether to access a document during a current crawl with the aid of a probabilistic model that is based on the probability that the document has changed since a previous crawl. The foregoing language has nothing whatsoever to do with probability, which, as noted above, is based on the ratio of the number of times an event occurs during a larger number of trials. The foregoing language states that determining how many new and old Web pages to evaluate is based on "the amount of old pages already listed in the crawl database." This has nothing whatsoever to do with determining the probability that a document has changed since a previous crawl. Consequently, applicants respectfully submit that Claim 1 is not anticipated by Chakrabarti et al. and, thus, is clearly allowable.

Applicants also submit that Claims 2-9, all of which depend directly or indirectly from Claim 1, are allowable for the same reasons that Claim 1 is allowable.

Applicants further submit that Claims 2-4 and 6, which were rejected on the same grounds as Claim 1, are allowable for additional reasons. For example, Claim 2, which depends from Claim 1, recites that determining whether to access a document with the aid of a probabilistic model comprises computing a probability that the document has changed since the document was created during the previous crawl. Applicants submit that **Chakrabarti et al. teaches no such computation at Col. 9, lines 56-63 (reproduced below in connection with Claim 10).** As a result, applicants respectfully submit that Claim 2 is allowable for reasons in addition to the reasons why Claim 1 is allowable.

Claim 3, which depends upon Claim 2, recites that computing a probability that the document has changed comprises: selecting an active probability indicative of a proportion of documents in a plurality of documents that are changing at various rates, the plurality of documents including the document; training the active probability to reflect experience with the document during a plurality of previous crawls; and using the trained active probability to compute the probability that the document has changed. Applicants respectfully submit that the subject matter of Claim 3, particularly when taken in combination with the subject matter of Claims 1 and 2, is clearly not taught or even remotely suggested by Chakrabarti et al.

Contrary to remarks set forth in the Office Action, indicating the date and time when a Web page was last modified does not teach or even remotely suggest selecting an active probability indicative of a portion of documents in a plurality of documents that are changing at various change rates, the plurality of documents including the document. As noted above, Chakrabarti et al. has nothing whatsoever to do with probabilities and, thus, clearly does not select an "active probability" indicative of a portion of documents in a plurality of documents that are changing at various change rates, the plurality of documents including the document. Nor does Chakrabarti et al.'s suggestion that the topic itself can be defined by a user or by considering a seed set using a topic analyzer, including an associated classifier trainer, teach or suggest training an active probability to reflect experience with a document during a plurality of previous crawls. Further, computing a checksum representative of a page content does not teach or even remotely suggest using a trained active probability to compute a probability that a document has changed. Thus, applicants respectfully submit that Claim 3 is clearly allowable for reasons in addition to the reasons why Claims 1 and 2 are allowable.

Claim 4 is dependent upon Claim 3 and recites that the method further comprises selecting the probability that the document has changed from the previous crawl as the active probability in the current crawl and repeating the method of Claim 3 for the current crawl. Again, applicants respectfully submit that this subject matter is not taught or even remotely suggested by Chakrabarti et al. More specifically, this subject matter is clearly not taught or even remotely suggested by Chakrabarti et al.'s provision of an indication of the date and time of when a Web page was last modified, as asserted in the remarks accompanying the rejection of Claim 4. Thus, applicants submit that Claim 4 is allowable for reasons in addition to the reasons why Claims 1-3 are allowable.

Claim 6 is dependent upon Claim 1 and recites that the probabilistic model further comprises: training a document probability distribution corresponding to the document address specification to reflect experience with the document during a plurality of previous crawls, the document probability distribution including a plurality of probabilities; determining from the document probability distribution a probability that the document has changed; and making a determination of whether to access the document in a current crawl based on the probability that the document has changed. This subject matter is also not taught or remotely suggested by Chakrabarti et al.

More specifically, the employment of an associated classified trainer, as suggested at Col. 6, lines 7-15, of Chakrabarti et al., does not teach or even remotely suggest training a document **probability distribution** corresponding to the document address specification **to reflect an experience with the document during a plurality of previous crawls**. Nor does the suggestion of associated classifier trainers suggest a document probability distribution, including a plurality of probabilities. Further, the language of Col. 8, lines 53-67, clearly does not teach or even remotely suggest determining from a document probability distribution a probability that the document has changed. Applicants submit that the wording "evaluate for potential changes in the old pages that might have occurred since last time the old pages were considered by the system" is taken out of context. That language must be considered in combination with the language "the numbers of new and old pages generated at block 70 depends on, e.g., the amount of old pages already listed in the database 30." This language clearly has nothing whatsoever to do with probabilities. Nor does Col. 9, lines 45-63, quoted below in connection with Claim 10, teach or suggest using a trained active probability to compute the probability that a document has changed. As a result, applicants respectfully submit that Claim 6 is allowable for reasons in addition to the reasons why Claim 1 is allowable.

As amended, independent Claim 10 reads as follows:

> 10. A computer-readable medium having computer-executable instructions for retrieving one document in a plurality of documents from a remote server, which when executed comprise:
>
> maintaining historical information associated with changes to the one document;
>
> initiating a crawl procedure for retrieving particular documents in the plurality of documents; and
>
> determining whether to access the one document from the remote server based on a probabilistic analysis of the historical information associated with the changes to the one document, said probabilistic analysis of the historical information being based on the probability that the one document has changed since a previous crawl.

Clearly, Chakrabarti et al. teaches nothing whatsoever regarding determining whether to access a document from a remote server based on a probabilistic analysis of historical information associated with changes to the document, where the probabilistic analysis of the historical information is based on the probability that a document has changed since a previous crawl. Remarks included in the Office Action suggest that determining whether to access one document from a remote server based on a probabilistic analysis of the historical information associated with **changes** to the one document is anticipated by Col. 7, lines 45-63, of Chakrabarti et al. Applicants respectfully disagree. Col. 7, lines 45-63, of Chakrabarti et al. reads as follows:

> Moving to decision diamond 90 the worker thread determines whether the assigned page is a new page or an old page. If the page is an old page the logic moves to block 92 **to retrieve only the modified portions, if any, of the page, i.e., the portions that the associated Web server indicates have changed since the last time the page was considered by the system 10.** Accordingly, at decision diamond 94 it is determined by the system 10 whether in fact the old

> page has been changed as reported by the associated Web server, and if the page has not been changed, the process loops back to the sleep state at block 86.
>
> In contrast, **if the page is an old page that has been determined to have changed at decision diamond 94,** or if the page is determined to be a new page at decision diamond 90, **the logic moves to block 96 to retrieve the entire page from the associated Web server.** At block 98, a checksum representative of the page's content is computed, and this checksum establishes the OID field 38 (FIG. 1) of the associated entry in the Web page table 32. (Emphasis added.)

The bolded language clearly does not teach or suggest determining whether to access a document from a remote server based on a probabilistic analysis of the historical information associated with changes to the document, the probabilistic analysis of the historical information based on the probability that the document has changed since a previous crawl. The quoted language has nothing whatsoever to do with probabilistic analysis of historical information associated with changes to a document, much less a probabilistic analysis that is based on the probability that a document has changed since a previous crawl. Rather, the bolded language relates to retrieving pages if a Web server indicates that the pages have changed since the last time a page was considered by the system. Chakrabarti et al. is based on whether or not a change has occurred. It is not based on the probability that a change has occurred. Consequently, applicants submit that Claim 10 is also clearly allowable.

Applicants also submit that all the claims that depend from Claim 10, i.e., Claims 11-20, are allowable for the same reasons that Claim 10 is allowable.

Applicants further submit that Claims 11-20 are allowable for additional reasons. Claim 11 recites that the instructions further comprise: if the determination to access the document (which is based on a probabilistic analysis of historical information per Claim 10) is positive, identifying the one document for retrieval during the crawl procedure; and attempting to retrieve all documents identified for retrieval during the crawl procedure. Claim 12, which depends from Claim 10, recites that the probabilistic analysis comprises computing a probability that the one document is changed since the one document was last retrieved from the remote server. As noted above, Chakrabarti et al. does not teach any form of probability computation. The calculation of a checksum based on a page's content does not anticipate computing a probability that a document has changed since the document was last retrieved.

Claim 13, which depends from Claim 12, recites that computing the probability that one document has changed further comprises beginning with a probability that a predefined portion of the documents in the plurality of documents has changed, training the probability that the predefined portion of documents has changed using historical information associated with the one document to achieve the probability that the one document has changed. Again, this subject matter is not taught or even remotely suggested by Chakrabarti et al.

Claim 14, which depends from Claim 12, recites that the instructions further comprise making a random decision to retrieve the one document wherein the random decision is based on the probability that the one document has changed. Claim 15 is dependent upon Claim 14 and recites that the random decision is further biased by a synchronization level configured to influence the random decision based on a predetermined degree of tolerance for not retrieving the one document that the document is likely to have changed. Claim 16 is dependent upon Claim 14 and recites that the random decision is made by a software routine adapted to simulate the flip of a coin.

Clearly, the subject matter of Claims 11-16 is not taught or suggested by Chakrabarti et al., particularly when the subject matter of these claims is considered in combination with the subject matter of the claims from which they depend. Thus, applicants submit that these claims are allowable for reasons in addition to the reasons why Claim 10 is allowable.

Claim 17 is dependent upon Claim 10 and recites that: the historical information associated with changes to the one document includes a time stamp for the one document, the time stamp being indicative of the time that the one document was last modified when the one document was last retrieved from the remote server; and the probabilistic analysis includes a comparison with the time stamp included in historical information with another time stamp associated with the one document stored on the remote server. Claim 18 is dependent upon Claim 17 and recites that if the time stamp included in the historical information does not match the other time stamp associated with the one document stored on the remote server, identifying the one document for retrieval during the crawl procedure. Again, applicants respectfully submit that the subject matter of Claims 17 and 18 is not even remotely suggested by Chakrabarti et al. and, thus, these claims are allowable for reasons in addition to the reasons why Claim 10 is allowable.

Claim 19 is dependent upon Claim 10 and recites that the historical information associated with changes to the one document includes a hash value associated with the one document, the hash value being a representation of the one document; and the probabilistic analysis includes a comparison of the hash value included in the historical information with another hash value calculated from information retrieved from the one document on the remote server. Claim 20 is dependent upon Claim 19 and recites that if the hash value included in the historical information does not match the other hash value associated with the one document stored in the remote server, identifying the one document for retrieval during the crawl procedure. Again, the subject matter of Claims 19 and 20, particularly when considered in combination with the subject matter of Claim 10, is clearly not taught or even remotely suggested by Chakrabarti et al. Thus, Claims 19 and 20 are also submitted to be allowable for reasons in addition to the reasons why Claim 10 is allowable.

Rejection of Claims 5 and 7-9

As noted above, Claims 5 and 7-9 were rejected under 35 U.S.C. § 103(a) as being unpatentable in view of the teachings of Chakrabarti et al. Applicants respectfully disagree. Remarks accompanying the rejection of Claim 5 recognize that Chakrabarti et al. does not expressly teach "training the active probability includes multiplying the active probability indicative of a change in the document of a trained probability calculated using a probabilistic model." The remarks state that Chakrabarti et al. suggests this subject matter at Col. 7, lines 3-17, which read as follows:

> To determine the relevance of a document, it is assumed that the category taxonomy imposes a hierarchical partition of web documents. Categories in the taxonomy tree, also referred to as nodes, are denoted "c". The predicate good(c) denotes whether a node "c" has been marked as good. By definition, **for any document "d", the probability that it was generated from the category c_0 corresponding to the root node, denoted Pr[c_0|d], is one.** In general Pr[c|d]=Pr[parent(c)|d]Pr[c|d,parent(c)]. Pr[parent(c)|d] is computer**[sic?]** recursively, whereas Pr[c|d,parent(c)] is computed using Bayes Rule as Pr[c|parent(c)]Pr[d|c/Σ_cPr[c'|parent(c')]Pr[d|c'], where the sum ranges over all siblings c' of c. Finally, **the probability that a page is relevant is $\Sigma_{good(C)}$Pr[c|d].** This quantity, denoted R(d), is typically very small, so the logarithm of R(d) can be stored if desired. (Emphasis added.)

Applicants submit that the bolded language clearly does not anticipate or even remotely suggest training an active probability by multiplying the active probability indicative of a change in the document by a training probability calculating using a probabilistic model. The probability determined by Chakrabarti et al. is the probability that a page is relevant, not the probability that a change has occurred in a document. Thus, applicants respectfully submit that Claim 5 is allowable for reasons in addition to reasons why the claims from which Claim 5 depends (Claims 1-3) are allowable.

Claim 7 is dependent upon Claim 6 and recites that the method further comprises: calculating, based on the experience of the document during a plurality of previous crawls, a discrete random variable distribution that includes a plurality of training probabilities; and multiplying each probability in the document probability distribution by a corresponding training probability from the discrete random variable distribution. As with the other dependent claims, this subject matter is not even remotely suggested by Chakrabarti et al.

Applicants respectfully submit that the Chakrabarti et al. language in Col. 7, lines 18-65 (stating that the priority of a document can not only be determined by determining its relevance, but also by determining its "popularity," a measure of the quality of the document) does not teach or even remotely suggest that subject matter of Claim 7. Thus, Claim 7 is submitted to be allowable for reasons in addition to the reasons why Claims 1 and 6 are submitted to be allowable.

Claim 8 is dependent on Claim 7 and recites that training probabilities are calculated using a Poisson process, the Poisson process including a particular Poisson equation and a complementary Poisson equation. Again, this subject matter is not taught or even remotely suggested by Chakrabarti et al.

The Office Action asserts that the subject matter of Claim 8 is suggested at Col. 7, lines 3-17, quoted above. Applicants respectfully disagree. There is simply no teaching of this subject matter, or even any remote suggestion thereof, in Col. 7, lines 3-17. Thus, applicants submit that Claim 8 is also allowable for reasons in addition to the reasons why the claims from which Claim 8 depends are allowable.

Claim 9 is dependent upon Claim 8 and recites that the experience with the document during the plurality of previous crawls is derived from historical information associated with the document address specification. While Chakrabarti et al. arguably does not employ historical information, applicants submit that when the subject matter of Claim 9 is considered in combination with the subject matter of the claims from which Claim 9 depends, the combination is clearly not taught or remotely suggested by Chakrabarti et al. Thus, Claim 9 is also submitted to be allowable.

In summary, dependent Claims 5 and 7-9 include additional recitations that further distinguish the claimed subject matter from the teachings of Chakrabarti et al. As a result, applicants respectfully submit that these claims are allowable for reasons in addition to the reasons why the claim (Claim 1) from which these claims directly or indirectly depend is allowable.

CONCLUSION

In view of the foregoing amendments and comments, applicants respectfully submit that all of the claims in this application are clearly not anticipated by Chakrabarti et al. and/or unpatentable in view of Chakrabarti et al and, thus, are allowable. Consequently, early and favorable action allowing these claims and passing this application to issue is respectfully solicited. If the Examiner has any questions, the Examiner is invited to contact applicants' attorney at the number set forth below.

Respectfully submitted,

CHRISTENSEN O'CONNOR
JOHNSON KINDNESS^{PLLC}

Gary S. Kindness
Gary S. Kindness
Registration No. 22,178
Direct Dial No. 206.695.1702

I hereby certify that this correspondence is being deposited with the U.S. Postal Service in a sealed envelope as first class mail with postage thereon fully prepaid and addressed to Mail Stop AF, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the below date.

Date: 10/7/04 Shannon Hill

GSK:snh